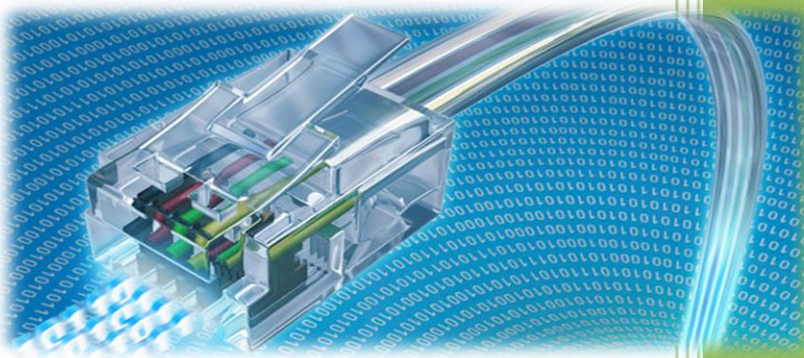


2013

Measuring Broadband America: Statistical Review and Advance Analysis



Practicum Team:

Attila Veres

Neha Rawal

Suman Basu

Robert Moreira

Institute for Advanced Analytics,
NCSU

4/10/2013

Table of Contents

TABLE OF CONTENTS	1
1 EXECUTIVE SUMMARY	3
1.1 PURPOSE.....	3
1.2 KEY FINDINGS.....	3
1.2.1 Data Usage and New Peak.....	3
1.2.2 Accounting for Variance	3
1.2.3 Important variables.....	3
1.2.4 Windows Application	4
1.3 RECOMMENDATIONS	4
2 STATISTICAL STUDIES	5
2.0.1 KEY FINDINGS.....	5
2.1 SAMPLING METHODOLOGY.....	7
2.1.1 Recommendations and Key Findings	7
2.1.2 Overview.....	7
2.1.3 Current Method	7
2.1.3.1 Cells.....	7
2.1.3.2 Sample Size.....	8
2.1.3.3 Volunteer Sample.....	8
2.1.3.4 Transparency	8
2.1.3.5 Collaboration.....	8
2.1.4 Suggestions	8
2.1.4.1 Embedded Modem Approach.....	8
2.2 STATISTICAL METHODOLOGY	10
2.2.1 Recommendations and Key Findings	10
2.2.2 Aggregation	10
2.2.3 Accounting for Variance	10
2.2.4 Formal Testing.....	10
2.2.5 Outliers.....	11
2.2.6 Confidence Intervals.....	11
2.2.6.1 Bootstrapping	11
2.2.7 Analysis of Data Usage	12
2.2.7.1 Peak period for data traffic.....	12
2.2.7.2 Time of the day effect.....	12
2.2.7.3 Weekend V/S Workday Effect.....	12
2.2.7.4 Other important factors	12
2.2.5 A Priori Hypotheses.....	13
2.3 ACCOUNTING FOR VARIANCE	14
2.3.1 Recommendations	14
2.3.2 Key Findings.....	14
2.3.3 Overview.....	14
2.3.4 The Consistent Speed Metric	14
2.3.5 Comparing Metrics.....	16
2.4 IMPORTANT VARIABLES.....	18

2.4.1 Key Findings.....	18
2.4.2 Conclusion.....	20
2.5 WINDOWS APPLICATION TO DISPLAY STATE-LEVEL ISP PERFORMANCE.....	21
2.5.1 Motivation	21
2.5.2 Application output	21
2.6 VISUALIZATIONS.....	23
2.6.1 Depicting Spatial Variance through Thematic Maps	23
2.6.2 Comparing Mean & Variance across ISPs through Error Bars.....	23
2.6.3 Comparing ISP presence in the different states.....	24
2.6.4 Comparing consistent actual / advertised speed across ISPs	25
2.7 VALIDATION AND REVIEW.....	26
2.7.1 April Analysis.....	26
2.7.1.1 Purpose	26
2.7.1.2 Recommendation	26
2.7.1.3 Finding.....	26
2.7.1.4 Deliverable	26
2.7.2 September Analysis.....	27
2.7.2.1 Purpose	27
2.7.2.2 Recommendations	27
2.7.2.3 Findings.....	27
2.7.2.4 Deliverable	27
2.8 FUTURE APPLICATION.....	28
2.8.1 Further Analyses	28
2.8.3 Mobile Program.....	28
4 TECHNICAL APPENDIX.....	29
4.1 TOOLS USED.....	29
4.2 APPENDIX A : DATA DICTIONARY	30
4.3 APPENDIX B: HISTOGRAMS	31
4.4 APPENDIX C : CONSISTENT SPEED METRIC.....	33
4.4.1 Appendix C1: Calculating the Metric.....	33
4.4.2 Confidence Intervals	34
4.5 APPENDIX C (VARIABLE IMPORTANCE).....	35
4.5.1 Appendix C1: Variable Selection Node in Enterprise Miner	35
4.5.2 Appendix C2: Decision Tree	36
4.5.3 Appendix C3: Linear Regression	36
4.5.4 Appendix C4: Ranking Methodology	37
4.6 APPENDIX D (BUILDING THE APP).....	40
4.6.1 System Requirements for the application	40
4.6.2 Preparing the data.....	40
4.6.3 Executing the application.....	40
4.6.3 Statistical considerations and drawbacks	43
4.7 APPENDIX D: VALIDATION AND REVIEW.....	44
4.7.1 Appendix D1: Methodology	44
4.7.2 Appendix D2: Examples of Validation Results	44
4.7.3 Appendix D3: Explanation of Bug	45

1 Executive Summary

1.1 Purpose

The Federal Communications Commission (FCC) started the Measuring Broadband America (MBA) study in March 2011. It has successfully run for the last three years, measuring and presenting the performance of residential broadband performance provided by different broadband service providers (ISPs). The goal of the IAA team was to make recommendations that could improve the existing fixed broadband program and be incorporated into the future wireless program as well.

1.2 Key Findings

1.2.1 Data Usage and New Peak

After examination of the data usage information provided by the FCC, it was discovered that the time at which the internet traffic is most congested (i.e. when data usage rates are highest) is 8pm-12am on Saturday and Sunday, not 7pm-11pm Monday through Friday. Peaks still occur on weekdays, but the weekend peak is far more pronounced. *Additional details can be found in section 2.2.7.*

1.2.2 Accounting for Variance

The previous method of reporting only averages does not adequately portray variance across time and across users, so the team created a new metric, consistent speed, to account for this deficiency. The metric looks at the 10th percentile of speed for each unit ID as well as the 10th percentile of speed for each ISP, effectively capturing 90% of users, 90% of the time. For instance, if an unit ID has a consistent speed of 80, then that user is getting at least 80% of their advertised speed 90% of the time. If an ISP has a consistent speed of 80, then that ISP provides 90% of its customers with at least 80% of their advertised speed 90% of the time. *Additional details can be found in section 2.2.3.*

1.2.3 Important variables

The factors that are most influential in determining the consistent speed a customer receives are as follows (*NOTE: variable x variable denotes an interaction*):

Rank	Variable	Score
1	Download Tier x ISP	76
2	Peak x ISP	75
3	Download Tier x State	68
4	Total Data	62
5	ISP x State	59
6	Download Tier x Validated By	57

7	Download Tier	53
8	ISP	53
9	Validated By	50
10	State x Technology	50
11	Peak	48
12	State	44
13	State x Validated By	39
14	ISP x State x Technology	38
15	Days	37
16	ISP x Validated By	32
17	Peak x Technology	26
18	Technology	20
19	ISP x Technology	20

For detailed information regarding the scoring techniques used to obtain this data, please refer to section 2.4.

1.2.4 Windows Application

If a picture says a thousand words, then an application says a million. In order to make the material contained in these reports more useful to the end consumer, it is suggested that the Windows application submitted with this report be deployed to give consumers information that is most relevant to them. The FCC can also convert the application to whatever format is deemed appropriate for easy consumption by the end consumer.

1.3 Recommendations

Many of the findings for the fixed program could also be applied to the mobile program that is underway.

- Use formal statistical tests to validate the inferences from the analysis.
- Use the consistent speed (10th percentile) metric to account for both temporal and spatial variance.
- Use visuals that can convey complex information in easy to understand forms.

2 Statistical Studies

2.0.1 Key Findings

The key aspects studied and key findings (KF) identified by the team were:

- 1. Sampling Methodology:** The sampling approach followed by the FCC in this study is based on measuring the performance of around 8,000 volunteers. The current approach was studied and compared to other potential approaches in section 2.1.
Key finding: *It was found that the stratified sampling approach by the FCC is good and effective given privacy and cost constraints.*
- 2. Statistical Methodology:** A study was done to understand the statistical effectiveness of the analysis. Section 2.2 explains some of the pros and cons of the current approach and also suggests improvements.
Key finding: *Formal statistical tests (bootstrapping to build confidence intervals) should be run to validate the inference from the analysis.*
- 3. Peak Time:** Analysis of data usage seems to indicate a different peak time than what was originally indicated as the time of maximum internet traffic, which was defined as 7pm to 11pm Monday through Friday. The new finding puts peak time between 8pm and midnight on Saturday and Sunday.
Key finding: *Either switch the definition of peak to reflect this new information or simply merge the two peak times to get a more accurate representation of peak.*
- 4. Accounting for Variance:** While the mean performance metric used in the existing study does a good job of giving the central tendency, it is not adequate to represent variance in broadband performance. The data collected in the study should account for two types of variance – a) variance in time (temporal variance) and b) variance across the unit IDs (spatial variance). Section 2.3 talks about how using the 10th percentile metric (consistent speed) accounts for both these variances.
Key Finding: *The consistent speed (10th percentile) metric helps paint a more realistic picture for the end consumer by accounting for variance and should be incorporated in future analyses.*
- 5. Important variables:** Several factors about a broadband connection can play a role in the performance observed. The influence of factors like – ISP, technology, download tier, peak time, total data and other variables as well as their interactions are studied in Section 2.4.

Key finding: *The significant and important factors that aid in determining consistent speed, including but not limited to ISP, peak, and State, are all listed in table 2.4.1.1. It was observed that all these factors together were only able to explain 40% of the variation in the consistent speed metric, which means there are other factors that contribute to broadband performance which are not being measured in the study as of yet. Perhaps demographic data can help bring the explanatory power of these models higher.*

6. **Validated By:** It was observed that VALIDATED BY was an important variable in predicting the consistent speed. It was noted that when the ISP validated the download speed tier the consistent speed would on average drop by 30% while when Sam Knows validated the download speed tier the consistent speed increased by 1%.

Key finding:

Thus relying on any one source as is done in most online broadband speed comparisons would introduce a large variation in speed tier estimate. It is advisable that the current FCC approach of validating the speed tier observed through testing the connection with the speed tier obtained from the ISP and the end consumer be followed in order to reduce such variation.

7. **Visualization:** The success of the MBA study lies in presenting the performance findings to the end consumer in an easily understandable form. Section 2.6 showcases new visuals created by the team that aim to communicate complex information in a simple and aesthetic manner.

Key finding:

- a. Charts that display variance in broadband performance.*
- b. Charts that display geographical heat maps giving information at a state-level .*
- c. The application explained in Section 2.5 displays performance based on the consumer's location (specifically, state in this case).*

8. **Validation & Review:** The results presented in the April 2012 and September 2012 report were independently validated by the team.

Key finding: *The results from the validation as discussed in section 2.7 were within a margin of error of 2% to those in the reports, thus validating the analysis done by the FCC.*

2.1 Sampling Methodology

2.1.1 Recommendations and Key Findings

- The total United States sample size is adequate for statistical inference at a national level.
- However, some states have extremely low sample sizes; therefore, no statistical inferences can be made about those locales.
- Regional inferences can be made depending on how the regions are split.
- Some ISPs did not have enough participants to adequately report their performance.
- Volunteer sample does not seem to introduce bias into the measurement results as currently deployed.
- Ideally, the FCC should consider adopting a new sampling approach where the measuring tools are embedded in modems and tested randomly.

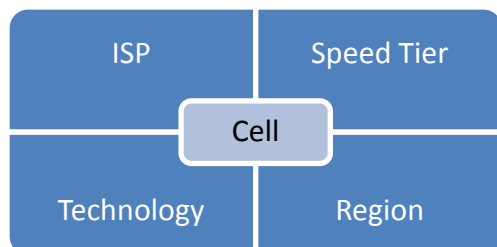
2.1.2 Overview

The goal of this section is to review previous findings and make final recommendations regarding the sampling methodology currently employed by the FCC for the MBA program. Overall, the FCC did a commendable job of creating a statistically valid, representative sample, especially given the financial cost of deploying such an ambitious endeavor. There is certainly room for improvement in the sampling methodology and suggestions are detailed below, but even if the FCC decides to continue with its current method, there is little concern about data integrity at this time.

2.1.3 Current Method

2.1.3.1 Cells

The method of splitting participants based on the key factors of ISP, region, speed tier, and technology is a great way to ensure representation across the United States. It is recommended that this remains a key part of dividing the sample as it has been successful thus far. The figure below summarizes the construction of each cell, stratified by ISP, Speed Tier, Technology and Region.



2.1.3.2 Sample Size

The current sample is large enough for statistical inference for the U.S. as a whole, but it is insufficient at a state or more granular level in most areas. In addition, a few ISPs were underrepresented in the sample and so no inferences can be made regarding these ISPs. Satellite technology was also of insufficient sample size to be considered in any analyses. Ultimately, given the restraint of using the white boxes and the associated cost of this method, it is unclear how to increase the ability to make local inferences without abandoning the white box method and switching to an embedded modem approach.

2.1.3.3 Volunteer Sample

Often times, the use of an all-volunteer sample might be a cause for concern given the tendency for such studies to introduce bias into the data. Volunteers often have specific motives when opening themselves up for study that make them different from the general population at large, making inferences to the greater population suspect at best. However, given the fact that the white box is completely independent of the end-user, there is no evidence that bias is an issue in the case of the MBA program. This is not to say the users who volunteer are not different from those in the population who do not volunteer, but since they have no effect on the technology that is measuring the speed, there should be no bias in the data currently being collected.

2.1.3.4 Transparency

The main benefit of the all-volunteer panel is that the end-user is well aware that he/she is being monitored and gives the program obvious transparency. In terms of replication of the report, it was also found that the majority of data and instructions were readily available on the FCC website. Originally there were a few missing pieces, but those have subsequently been added.

2.1.3.5 Collaboration

The current collaborative cooperation that the FCC has gained from ISPs in conducting this research has played an important role in getting the information required to perform valid analyses and is encouraged. The only area for concern in this regard is the possibility of an ISP intentionally boosting speed on a customer who is known to have a monitoring device on his/her line. However, as long as adequate safeguards are in place to monitor such an event, this sort of gaming is unlikely to occur.

2.1.4 Suggestions

2.1.4.1 Embedded Modem Approach

Given the ability to do so, the FCC should consider switching the sampling approach to a method where instead of white boxes, which are expensive to manufacture and distribute, the FCC simply embed the modems given by the ISP to the consumer with the technology to measure broadband speed in any home in which such a modem is deployed. This would

facilitate improvement of the current sampling methodology in 3 key ways; cost reduction, randomization, and sample size increase.

If the ISPs were responsible for deploying such boxes, it would shift the cost of manufacture and distribution from the FCC and would allow for funding to be allocated elsewhere. Given the current trend of fiscal austerity in government, reducing the cost of deploying a box to every home is certainly advisable.

In changing to this embedded method, the FCC would also obtain a much more randomized sample of the United States. Random samples are by far the most representative of a population because every single person in the population of interest has an equal chance of being selected to participate.

Finally, and most importantly, this method would allow the FCC to increase sample sizes in poorly represented localities. Given the reduced cost of employing such a method, the FCC could conceivably sample 100,000 boxes divided into the same representative cells and gain insights into states which are currently underrepresented, such as Alaska; the same method would also solve the issue of ISP underrepresentation.

2.2 Statistical Methodology

2.2.1 Recommendations and Key Findings

- Investigate the new peak time finding to determine whether original peak time estimates need to be updated.
- Report the variation that a customer can expect to receive from the ISP by using the consistent speed metric detailed in the next section or by some other means.
- Aggregation by Unit ID is the ideal way to reach maximum granularity while still maintaining independence in the data.
- Use confidence intervals, bootstrapped or otherwise, to back up visual representations in the charts and conclusions in the report.
- Formulate hypotheses before reviewing the data to ensure that no unintended biases arise.

2.2.2 Aggregation

The data consists of two elements, unit IDs and the hourly tests that measure broadband performance on each unit ID. The hourly tests on a unit ID are called measurement units while the unit IDs themselves are called experiment units in statistical language.

The hourly test results (measurement units) cannot be considered independent as they are run on the same unit ID. Since statistical tests require independent observations, analyses cannot be conducted on the tests themselves; therefore, aggregation is required at the experiment unit (unit ID) level. Aggregating on the unit ID is the ideal way to maintain maximum granularity while gaining the power for statistical inference.

2.2.3 Accounting for Variance

After aggregation, the metric by which to analyze the data must be decided. In previous analyses, the mean of actual / advertised speed was used as the primary metric on which to base findings. While averages are great for most types of analyses, in this case the mean provides only a static snapshot of performance, not a range that accounts for variance. The 10th percentile (consistent actual / advertised speed) that takes variance into account by looking at the distribution makes for a better metric.

2.2.4 Formal Testing

Formal statistical tests help to back up claims about the data made from the basic summary statistics. In this case, a confidence interval around the median (50th percentile) is sufficient to back up the claims made about ISP performance. However, with the new metric creating a heavily left skewed distribution, bootstrapping was performed to normalize the distribution and to calculate robust confidence intervals. This bootstrapping and building of confidence intervals is discussed in detail in section 2.2.6.

2.2.5 Outliers

Outliers can create problems for any analysis. However, in this case, the outliers, such as those tests that exceeded the advertised speed contain valuable information about a connection and the ISP. For that purpose, it is recommended to keep outliers in the data (barring missing values, which should be purged or imputed) and simply reassign them to extreme values of the desired range; 0% or 100% of advertised speed in this case. This way, the outliers are still included but the data is now in a manageable range that is more conducive to statistical inference.

2.2.6 Confidence Intervals

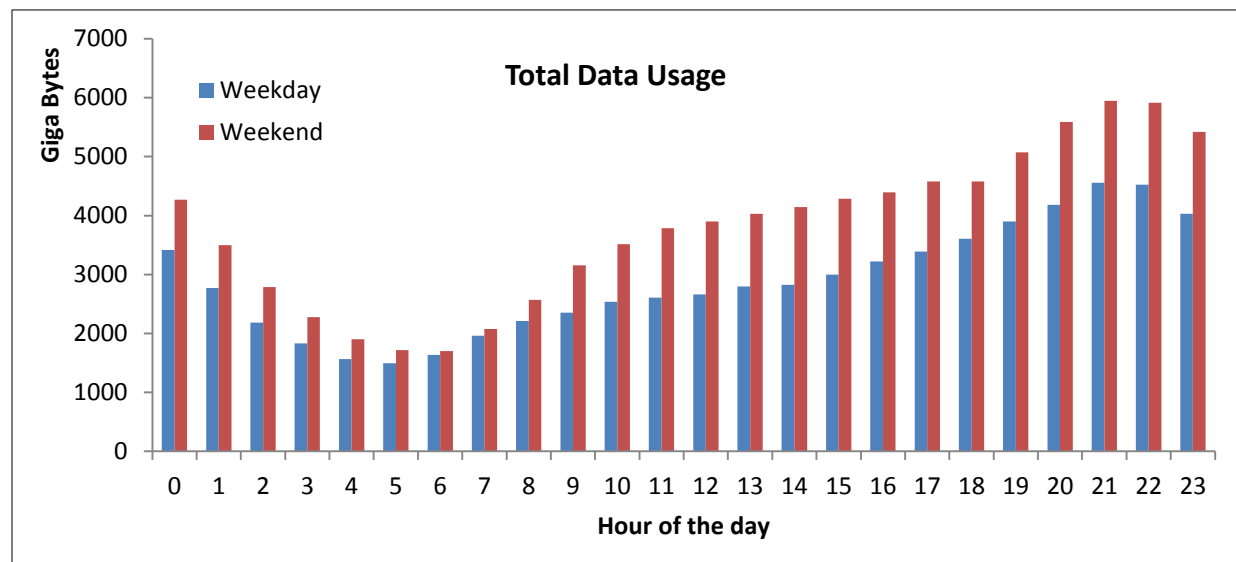
The next step in the process was to create a confidence interval around this new metric. However, the new consistent speed metric follows a non-normal distribution. Due to the consistent speed metric being bounded by 100, the distribution is highly left skewed, with most ISPs performing fairly well and tapering off to 3% of advertised speed at the tail. Confidence intervals require an assumption of normality in order to calculate the proper estimates, so bootstrapping had to be performed to generate a normal distribution from the values. The figure below is the confidence interval calculated without accounting for peak time for California only (all states are in Appendix C).

Figure 2.2.6.1: 99.99% Confidence Intervals for Median Consistent Speed				
ISP	State	Median	99.99% Confidence Interval	
ISP 1	CA	80	78	81
ISP 2	CA	83	83	83
ISP 3	CA	96	96	96
ISP 4	CA	99	99	100
ISP 5	CA	100	100	100
ISP 6	CA	91	76.5	100
ISP 7	CA	90.5	81	98.5
ISP 8	CA	98	98	98
ISP 9	CA	100	100	100
ISP 10	CA	100	97	100
ISP 11	CA	84	66	90

2.2.6.1 Bootstrapping

Bootstrapping is a data mining technique that draws multiple samples from the original sample to create a larger dataset that is more conducive to creating confidence intervals. By relying on the central limit theorem, the bootstrap will resample all of the data and create a normal distribution around each unit ID's consistent speed metric. After obtaining the bootstrapped sample, it is simple to calculate a confidence interval around the consistent speed metric.

2.2.7 Analysis of Data Usage



2.2.7.1 Peak period for data traffic

Comparing different hours of the day and different days of the week, it was noted that the peak data traffic time was 8pm – midnight on Saturday and Sunday.

2.2.7.2 Time of the day effect

- Maximum data used between 9 PM – 10 PM
- Minimum data used between 5 AM – 6 AM
- On average the data consumed at 9 PM was 32% more than that at 5 AM

2.2.7.3 Weekend V/S Workday Effect

On Weekends the users consumed 30% more data than on workdays. This percentage difference was dependent on the time of the day; while maximum difference was observed during the working hours 9 AM to 5PM; broadband usage was the same in the morning hours between 5 AM and 7 AM.

2.2.7.4 Other important factors

The effect of other factors influencing data consumption was studied but no concrete results were observed. Several models including stepwise, forward, and backward linear regression as well as decision tree models were run. Variables used to try to predict data usage included consistent speed, download tier, ISP, peak, validated by, state, days, technology, and all possible two way interactions of the class variables. Each linear regression model, as well as the decision tree models had average squared errors in the magnitude of the millions, which indicated extremely poor models. The linear regression models were only able to explain about 10% of the variability in the data usage variable,

which also indicates extremely poor models. Given the existing variables in the study, they do not do a good job of predicting data usage.

2.2.5 A Priori Hypotheses

One important factor to consider is that any hypothesis testing that seeks to compare ISPs or performance metrics require hypotheses to be formulated before the data is examined in order to avoid bias. For this case, a bootstrapping procedure was used that alleviates the potential pitfalls that can arise when performing hypothesis tests on previously viewed data, but in the future it would be beneficial to form any and all hypotheses ahead of opening the dataset. It is also recommended to partition the data into training and validation datasets if any model building is to be performed.

2.3 Accounting for Variance

2.3.1 Recommendations

- By using the consistent speed metric, the FCC can represent variance in static charts because the metric takes consistency (and inversely variation) into account.
- Using the new metric will also reduce the problem of some ISPs scaling above 100% on average as this metric is bounded at 100%.

2.3.2 Key Findings

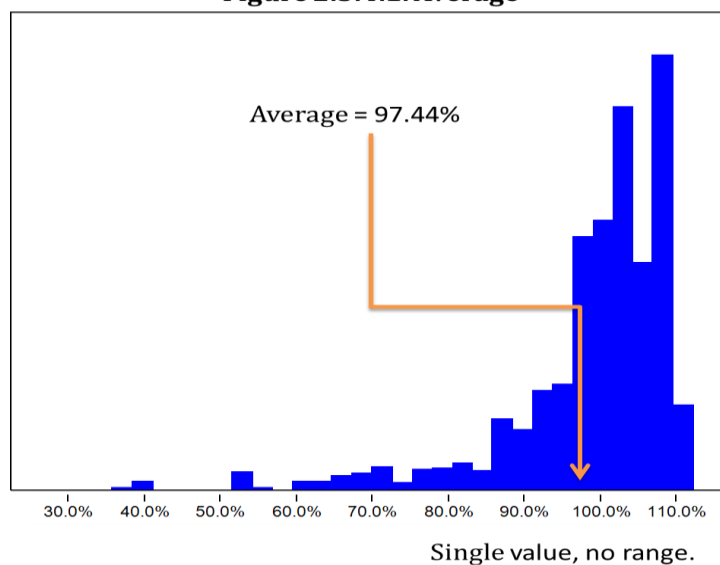
- Even though the new consistent speed metric penalizes ISPs who do not maintain a high speed consistently, most ISPs still do rather well overall.
- The metric more accurately represents what percentage of advertised speed a user will likely get from a given ISP.

2.3.3 Overview

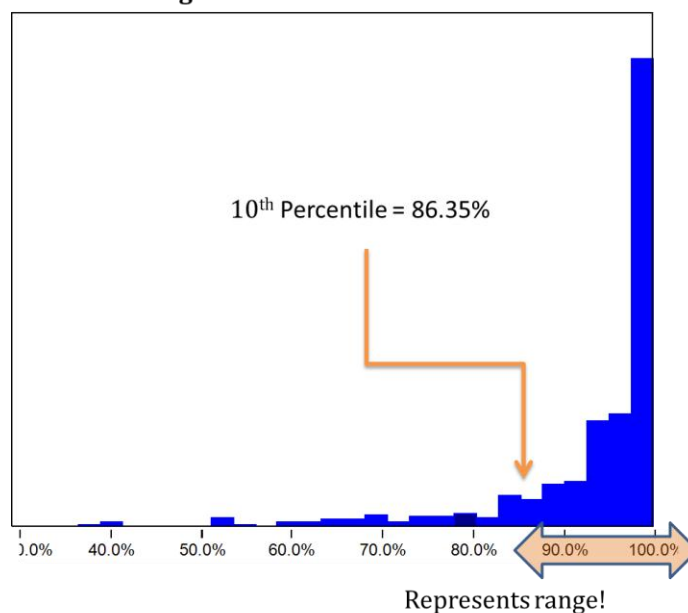
The current statistic calculated to measure speed is the aggregated average of actual vs. advertised speed by ISP or technology. This method, while statistically valid, is insufficient to represent key aspects of the data, such as consistency and variation in speed. One way to solve this problem is to include various additional charts that show variance, but the additional graphics may be cumbersome to navigate and difficult for consumers to digest. The recommended way to solve the above problem is to simply create a new metric called “Consistent Speed” that takes into account the variance of the measured speed.

2.3.4 The Consistent Speed Metric

The metric, in essence, represents the 10th percentile of the measured speed for all tests per unit ID. For instance, if a user has a consistent speed of 75% of advertised speed, this means that all tests measured for that unit ID were at or above 75% of advertised speed at least 90% of the time (*For more specifics and information on how to calculate this metric, please refer to Appendix C*). The current method of averaging uses a single fixed point to represent the wide range of values experienced by users (*Figure 2.3.4.1 below*).

Figure 2.3.4.1: Average

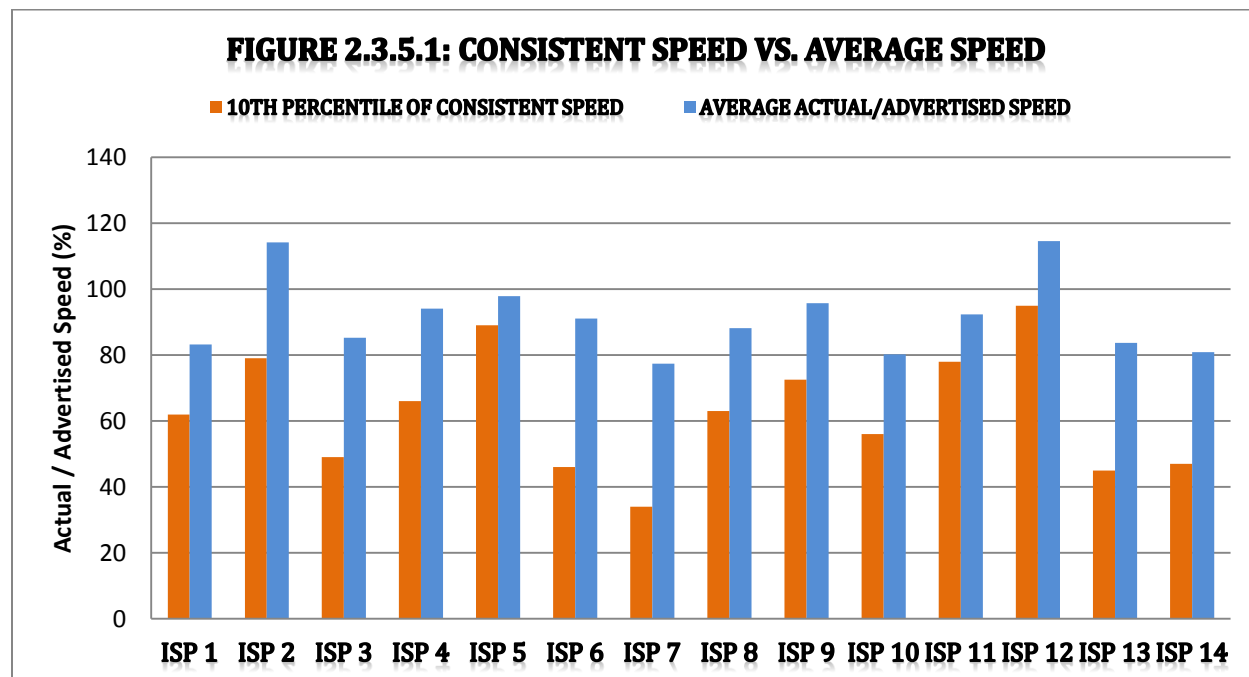
The above graph illustrates the aforementioned problem of representing only a single value in the current report, the average. Many users experience speeds above as well as below this value and the average is not representative of these users. However, by looking at the 10th percentile, a different story emerges (*Figure 2.3.4.2* below).

Figure 2.3.4.2: 10th Percentile

By looking at the 10th percentile, it is apparent that while the point value is lower (86.35% as opposed to 97.44%), it now represents 90% of all users. It is evident from the chart above that 90% of users are receiving at least 86.35% of advertised speed.

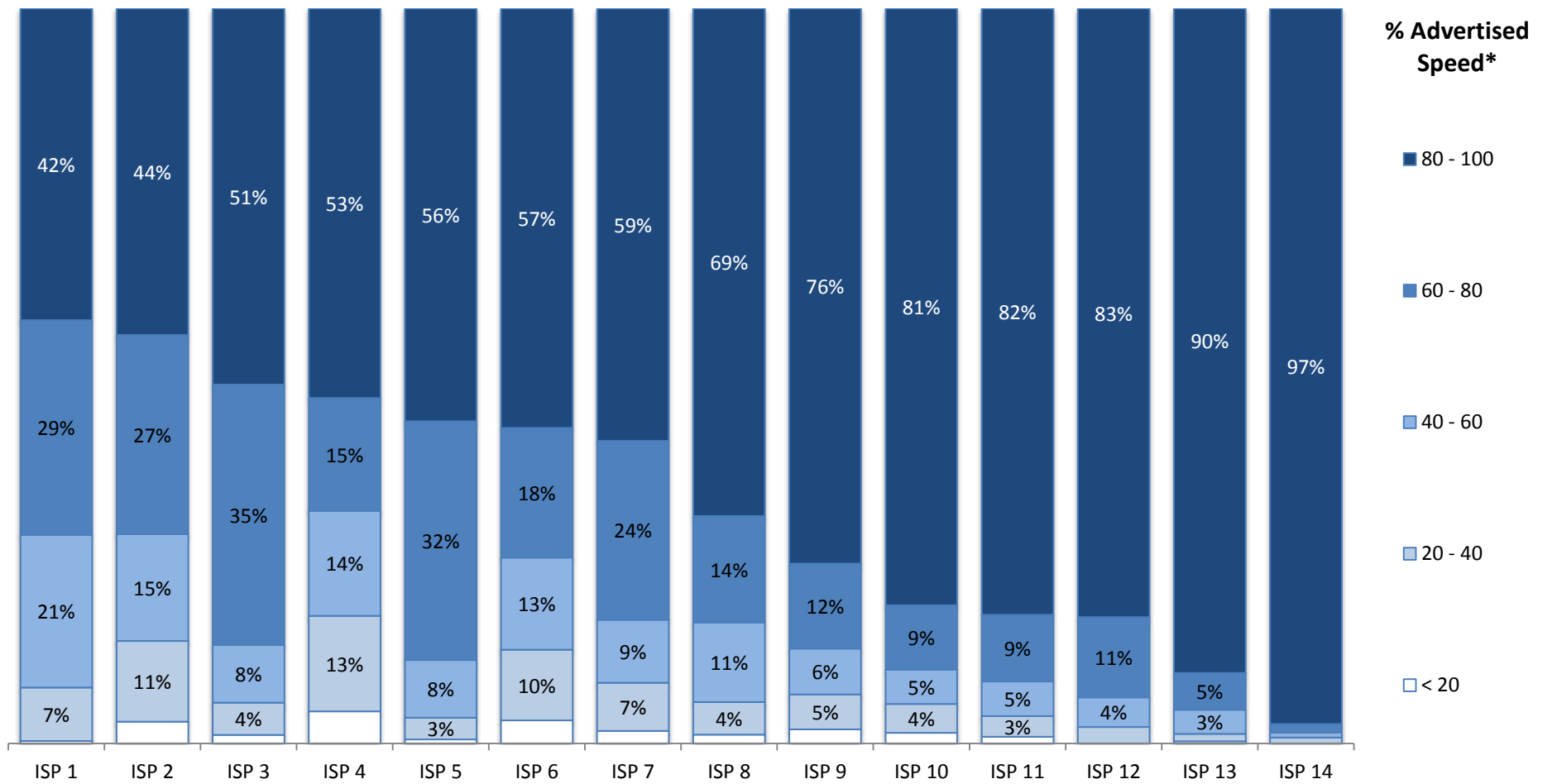
2.3.5 Comparing Metrics

The new consistent speed metric differs from the actual/advertised speed in a few key ways. The most obvious difference is the bounding of the metric to 100% whereas the original metric can exceed 100%. In addition, variation across unit IDs and across time does not show up in an average, but because consistent speed looks at the spatial and temporal variation in its calculation, variance is incorporated into the metric.



As evidenced above, the new metric also enables the static charts to account for variation in speed for a given user. When using an arithmetic mean, the variance disappears, but because this new metric looks at the percentage of time a connection goes beyond a certain threshold of speed, the more consistent connections get a higher consistent speed metric.

Distribution of Actual / Advertised Consistent Speed across ISPs



* 90% of the times

2.4 Important Variables

The goal of this section is to help the FCC understand which variables are most important in explaining what makes for a consistent user experience in regards to broadband service provider advertised speed. By getting a better understanding of which factors most explain a consistent broadband speed experience, the FCC can provide targeted advice to policy makers and the telecommunications industry on what factors to focus on in order to improve the general broadband experience in the United States. For example, in previous Measuring Broadband America reports there has been a focus on comparing broadband service providers alone to determine who provides a more consistent broadband speed experience for their customers. This study has found that looking at the broadband service provider alone does not explain who gets a consistent speed. However, looking at the broadband service provider interacting with other variables such as the time of day for example does help explain who gets a consistent broadband speed experience.

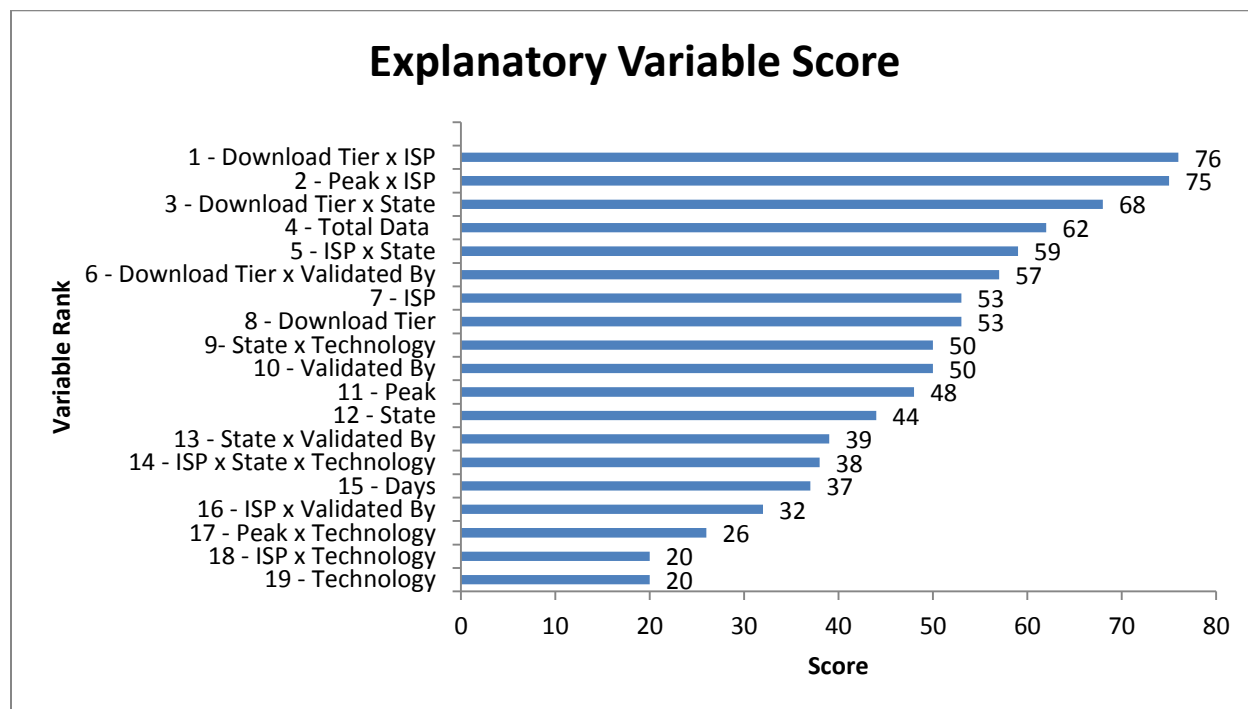
2.4.1 Key Findings

Different statistical variable selection techniques that leverage statistical significance such as p-value or explained variance such as R-square were used to select the best variables that explain consistent broadband speed. The primary focus of leveraging statistical techniques was to select variables and not to interpret inferences from parameter estimates. Many of the created interaction variables contain up to hundreds of levels per variable and therefore would make it difficult to gain inferences from a class variable. For detailed information on class level parameters, output for grouped levels per class variable from the sequential R-square selection technique has been provided in the Soft Appendix as an excel file (Variable Selection Grouped Levels.xlsx). After evaluating the variables each technique selected, a scoring system was developed in order to rank the most important explanatory variables. The highest possible score for a variable was 80 points where the lowest possible score was one point. The result of the selection techniques and weighting system are below:

Table 2.4.1.1: List of Important Variables that Influence Internet Performance

Rank	Variable	Score
1	Download Tier x ISP	76
2	Peak x ISP	75
3	Download Tier x State	68
4	Total Data	62
5	ISP x State	59
6	Download Tier x Validated By	57
7	Download Tier	53
8	ISP	53
9	Validated By	50

10	State x Technology	50
11	Peak	48
12	State	44
13	State x Validated By	39
14	ISP x State x Technology	38
15	Days	37
16	ISP x Validated By	32
17	Peak x Technology	26
18	Technology	20
19	ISP x Technology	20



The inference that can be gained from an interaction variable such as Download Tier x ISP implies that there is more explanatory power in looking at Download Tier and ISP and how their relationship affects the predicted variable Consistent Speed more powerfully than just looking at Download Tier and ISP separately. If one were to look at the highest ranking interaction variables, Download Tier x ISP, Peak x ISP, Download Tier x State, and ISP x State, one can infer that not only are these interaction variables important in statistically explaining more variability in Consistent Speed, but that their main effects variables, Download Tier, ISP, Peak, and State, also have explanatory significance.

The creation of interaction variables as well as new variables such as Peak, Total Data, and User Data helped in improving the explained variability in the Consistent Speed variable. In future studies more variables should be considered that can better explain Consistent Speed.

2.4.2 Conclusion

The models that were used in this study to determine which variables best explain Consistent Speed were able to explain 40% of Consistent Speed's variability. Although this is good, there is room for improvement. More or different types of variables can help in explaining more variability in the Consistent Speed variable. Certain variables that were created by the study such as Peak and Total Data certainly helped in further explaining the variability in Consistent Speed. Creating interaction variables also greatly increased the explanatory power of the models. The highest ranking variables were interaction variables that were created by linear regression selection techniques. Many of the same interaction variables also appeared in decision tree and sequential R-square selection techniques which showed that the interaction variables were significant across multiple techniques. The inference that can be gained from an interaction variable such as Download Tier x ISP is that looking at the relationship of those two variables and how they affect Consistent Speed has more explanatory power than looking just at Download Tier or ISP alone. The main purpose of the selection techniques was to select the most important variables to focus on to improve Broadband consistent speed. Because many of the variables had hundreds of levels, parameter estimation would have to be done on a one by one basis dependent on the exact level one wanted to focus on.

2.5 Windows Application to Display State-level ISP Performance

This section summarizes the IAA team's in-house developed Windows application that helps users to compare the ISP performances in their states.

2.5.1 Motivation

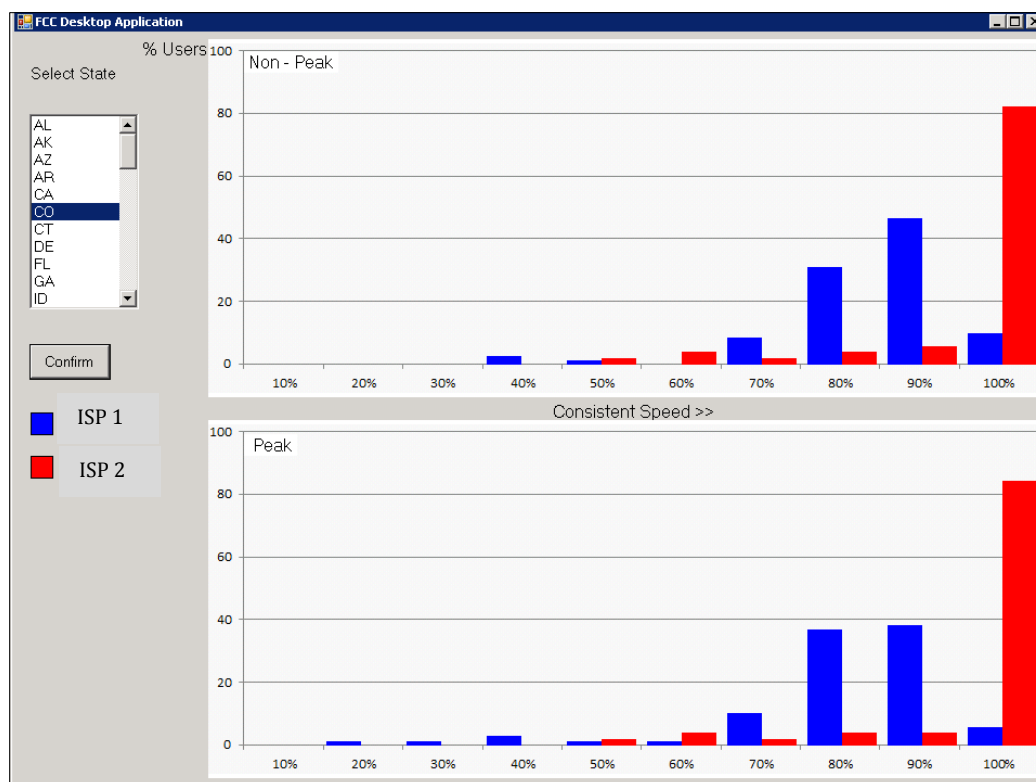
The motivation behind the application was to provide the broadband users with a specialized tool to assess their ISP's performance in their state using the consistent speed metric. This tool allows the broadband user to visually compare the broadband performance in his/her state which is more relevant information compared to the national averages.

The application lets the user select his/her state and displays distribution of the consistent speed metric (during peak hours and non-peak hours) for the top 4 ISPs (by number of unit IDs) in the state.

2.5.2 Application output

Once the database is prepared using the SAS Code (See Technical Appendix 4.6) the user can use the application to view the performance of the top 4 ISPs (or lesser), ranked by the number of unit IDs in his/her state.

The first bar chart shows the distribution of the consistent speed metrics in Non-Peak hours, while the second chart shows the same for the peak hours in a day. The X-axis represents the consistent speed while the Y-axis represents the percentage of total users in the state who experience that level of consistency.



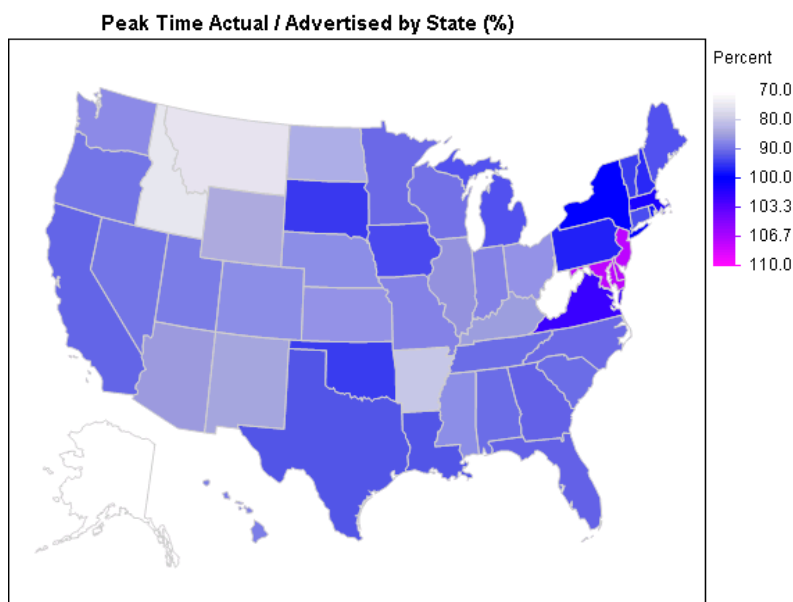
For example, in the above charts, about 45% of ISP 1 users in Colorado experience a consistent speed between 80% and 90% during non-peak hours, where as that proportion reduces to approximately 38% during peak hours.

2.6 Visualizations

This section showcases the different visuals that are useful to convey complex information in an easily understandable form.

2.6.1 Depicting Spatial Variance through Thematic Maps

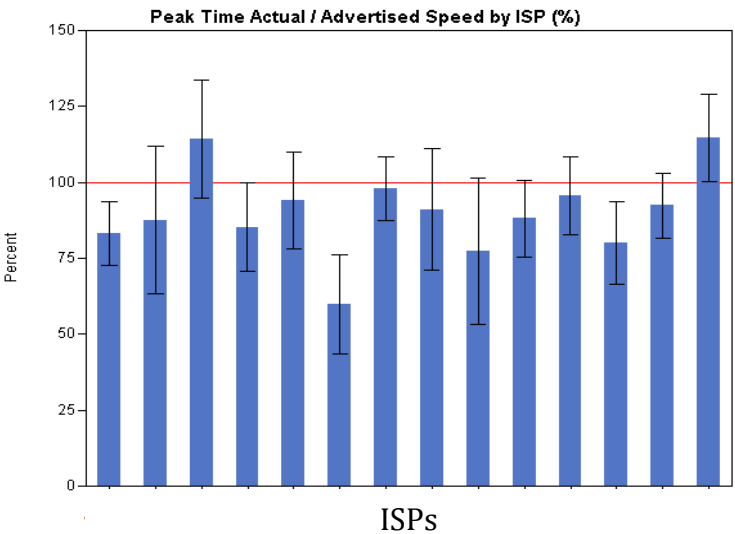
Before deciding to build an application, the IAA team attempted to depict variance by looking at a spatial thematic map, such as the one picture below.



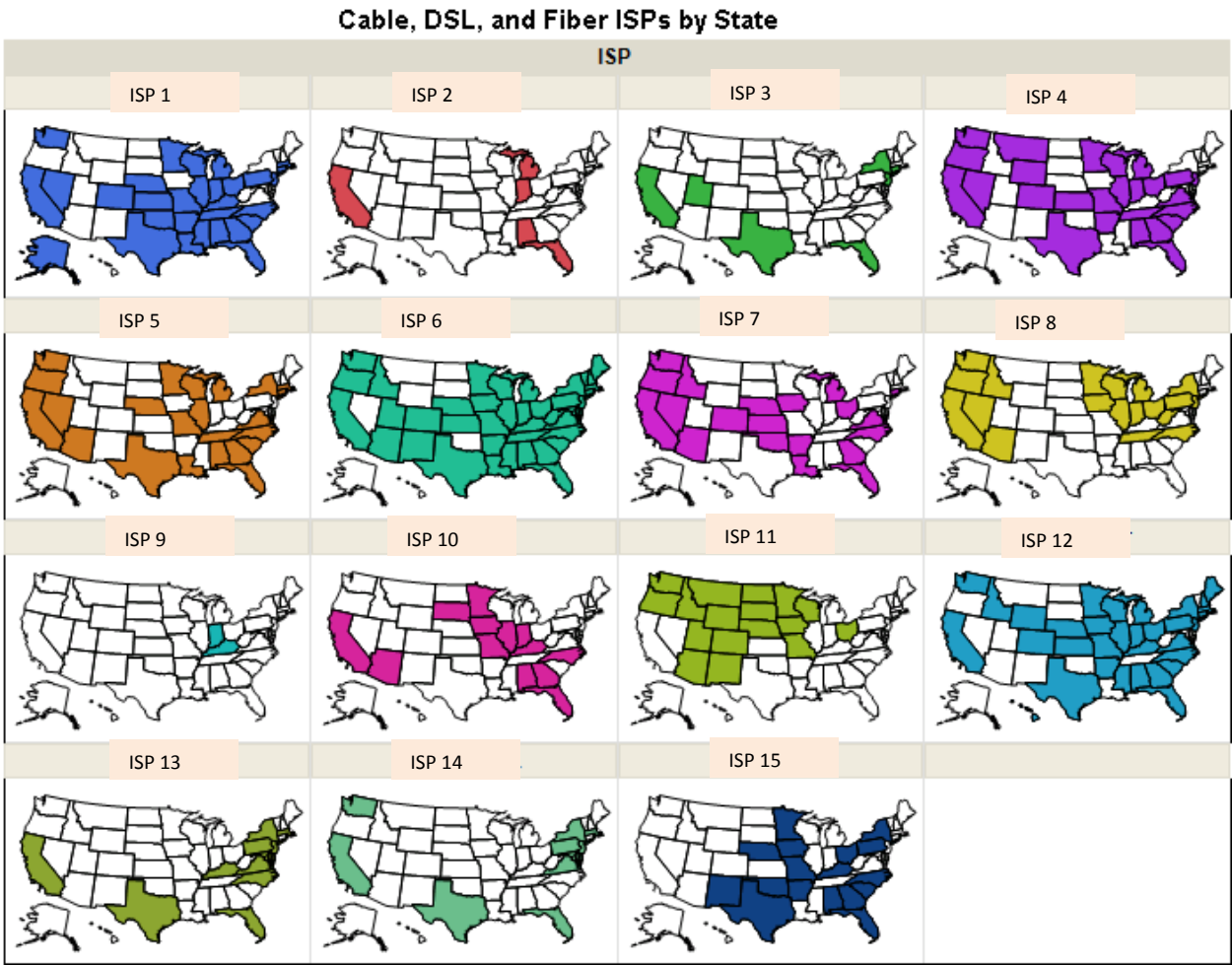
Ultimately, these representations proved useful but limited due to the static nature of the maps themselves and the variance in performance often found within states.

2.6.2 Comparing Mean & Variance across ISPs through Error Bars

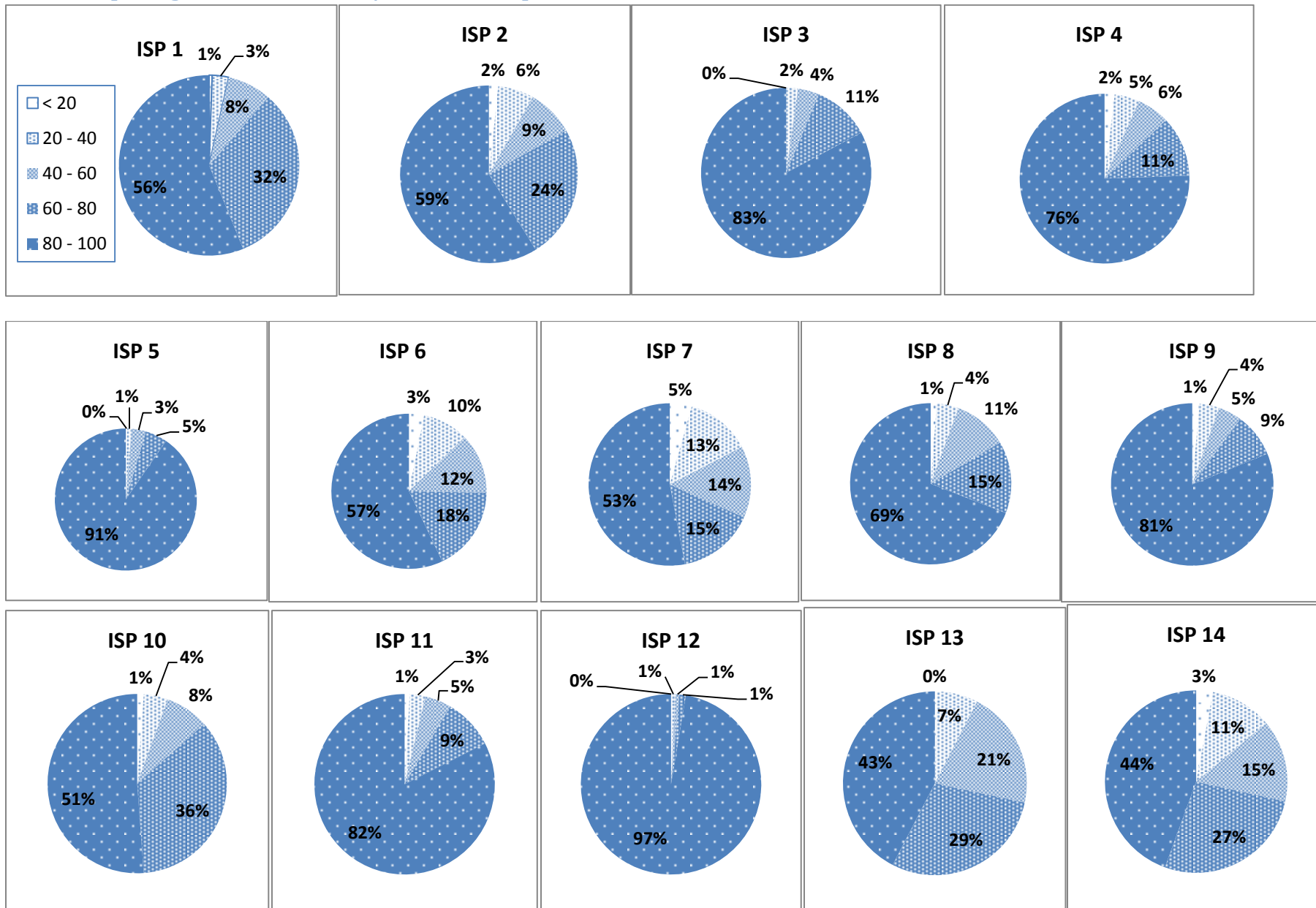
Another previously suggested idea was to plot a bar chart with whiskers indicating one standard deviation from the mean, such as the one below. Again, the graph is marginally useful, but it is difficult for the average consumer to interpret and while it picks up spatial variance, this sort of representation fails to account for temporal variance (example below)



2.6.3 Comparing ISP presence in the different states



2.6.4 Comparing consistent actual / advertised speed across ISPs



2.7 Validation and Review

The FCC has successfully run the Measuring Broadband America (MBA) initiative since 2011. This section is an overview of results of an independent validation of the MBA results by the IAA team. The review brought out new insights improving transparency of the process.

2.7.1 April Analysis

2.7.1.1 Purpose

The results from the April 2012 report were independently validated. Two tasks were undertaken:

- Running the existing scripts in order to generate results in the report
- Developing new scripts to validate the logic behind generation of results

[The detailed methodology is described in Appendix 4.7.1]

2.7.1.2 Recommendation

- Documentation of the current scripts needs to be improved by well commenting the code and using a data flow diagram.
- Making use of one tool for analysis is recommended, this makes understanding and validation of the analysis easier by any reviewer.

2.7.1.3 Finding

- Scripts, data and documents required for the validation, but missing for the FCC website were identified and reported.
- On average, a difference of 2% was observed between the existing results and those computed using the new SAS code.
[Examples listed in Appendix 4.7.2]
- A bug was observed in the existing SQL script and corrected for in the new SAS code.
[Explanation of bug correction provided in Appendix 4.7.3]

2.7.1.4 Deliverable

- The SAS codes were developed for validation of the following tests:
 - Netusage
 - Latency

[The codes and excel analysis are available in the Soft Appendix]

2.7.2 September Analysis

2.7.2.1 Purpose

The purpose of this section is to revisit the review for the Measuring Broadband America report for the September 2012 data. The IAA team has replicated all but 5 charts from the Statistical Averages dataset provided by the FCC. Five charts could not be replicated given the data provided because it would have required the non-aggregated raw data, which was not provided for the review.

2.7.2.2 Recommendations

Since a few of the charts could not be replicated, it is recommended for future reviews to include the following data:

- Unit level data
- Time point data
- Percentiles

2.7.2.3 Findings

- With the exception of those charts which could not be replicated due to insufficient data, all tables and charts that are a part of the upcoming Measuring Broadband America report were replicated to within an acceptable margin of error.
- For the charts that did contain some error, most of these can be easily attributed to difference in decimal places used in the original calculated tables and the ones created for this review.
- Only chart 22 contained a significant error, which can likely be attributed to the dropping of a few outlier variables in the original calculation as no outliers were dropped for this analysis.

2.7.2.4 Deliverable

The results of the analysis are available in the soft appendix in an excel file called Sept_AnalysisV1.0_Encrypted.xlsx

2.8 Future Application

2.8.1 Further Analyses

It is recommended to further explore the peak time and establish whether the primary peak of broadband usage is Monday through Friday 7pm-11pm or Saturday and Sunday between 8pm and midnight; or perhaps both. In the future, it is also a good idea to continue to use confidence intervals or hypothesis tests of some sort to establish a statistical backing to the claims made. In addition, given the various types of visualizations available in the market today, it would be beneficial to explore new avenues on that aspect. Finally, the application demonstrated by the IAA team should be adapted to fit the FCC website as a tool for consumers.

2.8.3 Mobile Program

Most of the analyses conducted for this report can be easily replicated for the mobile program by changing a few variable names in the code supplied with this report. The benefit of the mobile program will be primarily in the sample size issue present in this current data where local inferences are impossible due to low sample sizes. Therefore, some of the techniques, such as bootstrapping, may not be necessary to make inferences locally.

4 Technical Appendix

4.1 Tools Used

The following software was used to conduct all analyses, reports, and visualizations herein:

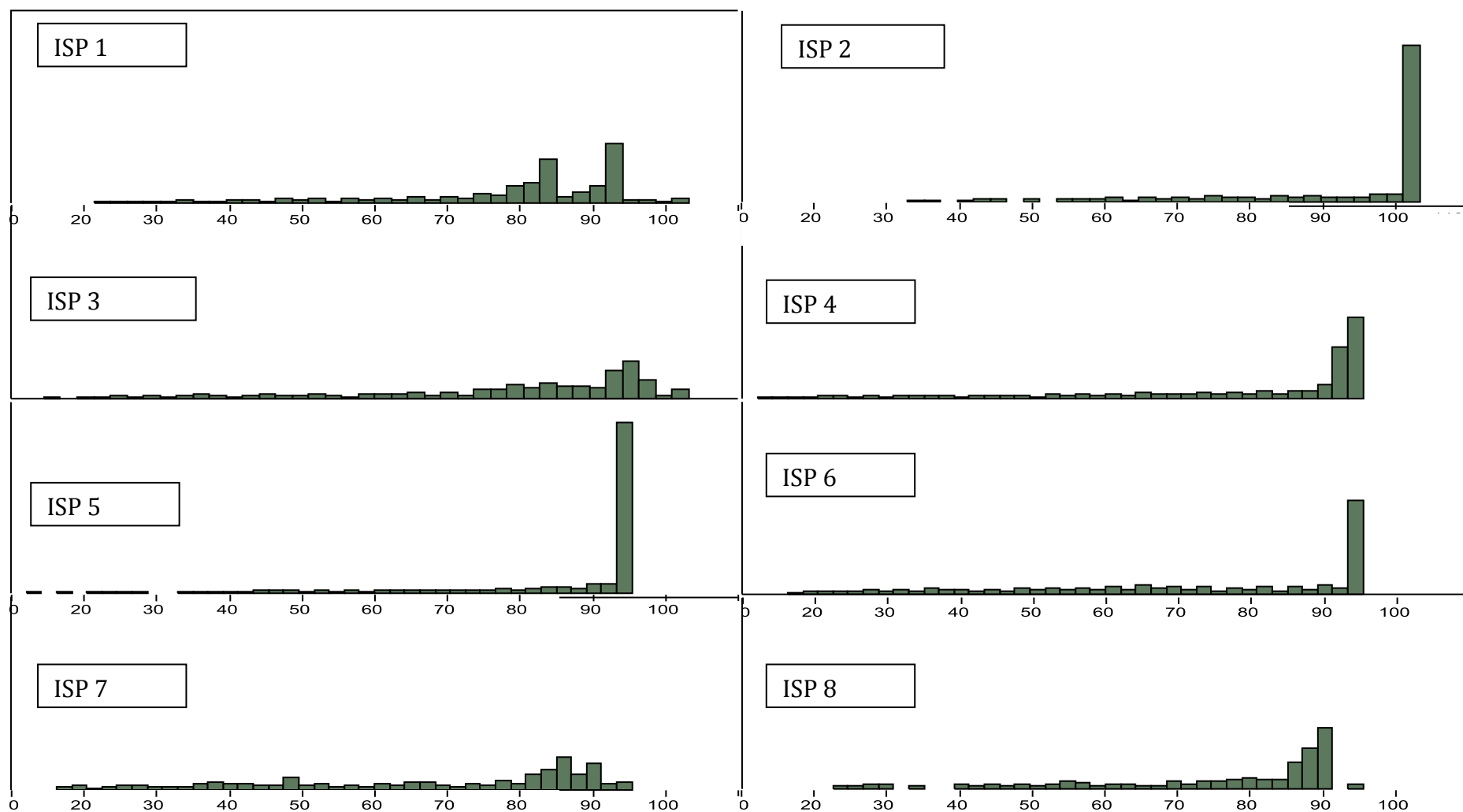
- SAS Version 9.3
- SAS Enterprise Guide version 5.1
- SAS Enterprise Miner Version 7.1
- JMP Pro 9
- Microsoft Office Professional 2010
- My SQL Server 5.5
- .NET Framework 3.5
- Visual Basic Power Packs

4.2 Appendix A : Data Dictionary

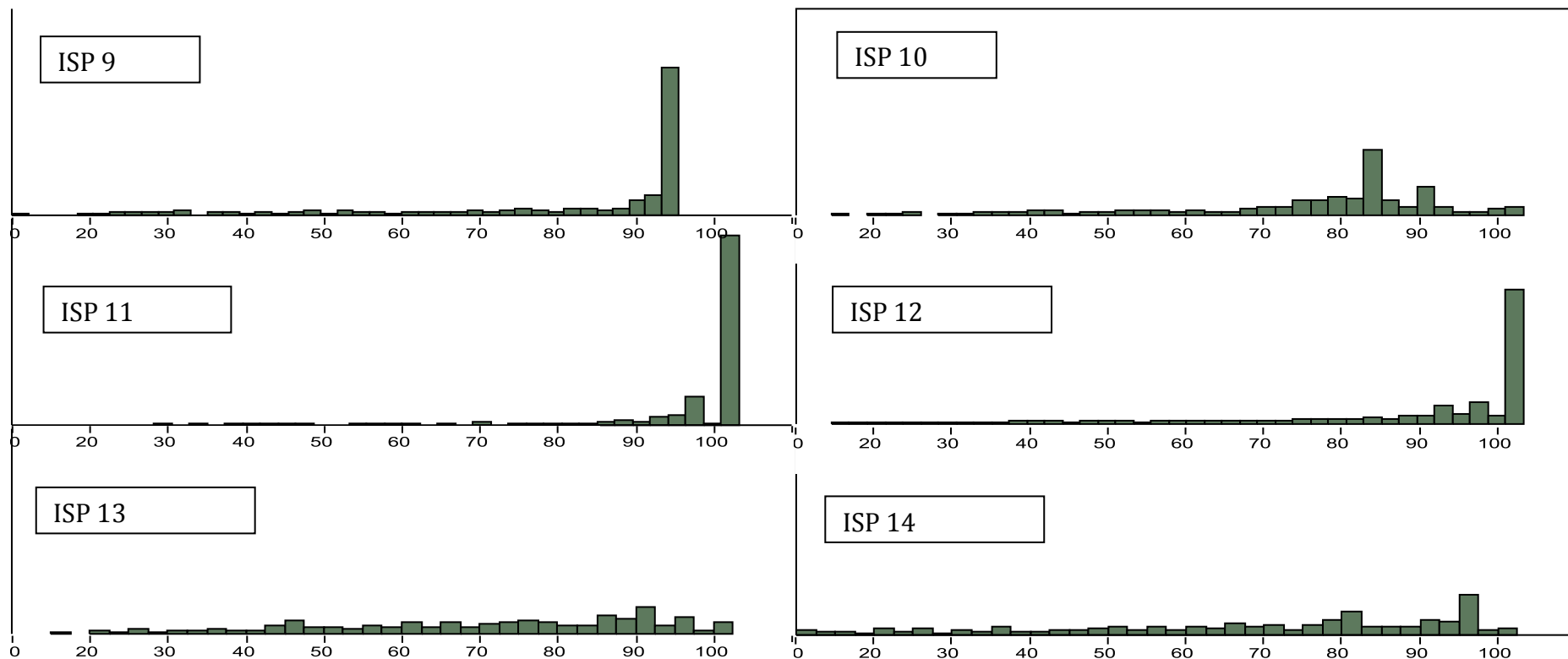
Consistent Speed Data Dictionary	
Consistent Speed	Interval variable that is bounded below by zero and above by 100 that measures the percentage of time a user attains their advertised speed or above. This is the target variable
Unit ID	Unique identifying number of network unit being observed.
Peak	Binary variable indicating if the consistent speed measurement was taken during peak or non-peak hours
ISP	A class variable indicating the Internet Service Provider of the observed unit
Download Tier	Ordinal variable indicating the ISP download tier of the observed unit
Technology	Class variable indicating the network technology of the observed unit
State	The US state where the observed unit resides
Region	Class variable that indicates the US region where the observed unit resides
Days	Number of days the unit was measured during the observed month of the study
Total Data	Data consumed by the user and MBA tests
User Data	Data Consumed by the user only
Mbps	Peak period average speed observed in Mbps
FCC Performance	Mbps/Download Tier %
Validated By	Class variable indicating whether the ISP, SamKnows, or both parties validated the download tier of the observed unit
ISP x State	Interaction class variable of ISP and State
ISP x Technology	Interaction class variable of ISP and Technology
ISP x Validated By	Interaction class variable of ISP and Validated By
ISP x State x Technology	Three way interaction variable of ISP, State, and Technology
Technology x State	Interaction variable of Technology and State
Peak x Technology	Interaction variable of Peak time and Technology
Peak x ISP	Interaction variable of Peak time and ISP
Peak x Download Tier	Interaction variable of Peak time and Download Tier

4.3 Appendix B: Histograms

Distribution of Consistent Speed for Each ISP



Distribution of Consistent Speed for Each ISP



4.4 Appendix C : Consistent Speed Metric

4.4.1 Appendix C1: Calculating the Metric

This consistent speed metric is calculated in the following way. The speed at a user's location is measured at various times of the day in bytes per second (bytes_sec), which is translated into megabits per second (Mbps) by dividing the bytes_sec variable by 131,072. Now the Mbps is known for each test run and can be compared to the advertised Mbps.

Once the measured and advertised speeds are on the same scale, a performance metric can be calculated for every single test. An array was created that contained 99 binary variables called Performance1-Performance100. The number at the tail end of the variable name represents a threshold by which the measured speed is compared to the advertised. For instance, Performance100 captures any test that achieved 100% or more of advertised speed, Performance99 captures any test that performed at .99*advertised speed, and so on until 1% of advertised speed. This variable gave every single test a score of 1 or 0 depending on whether or not the measured speed surpassed the specified threshold for the performance variable.

These performance metrics were then aggregated by frequency count on a unit ID level, which gave a cumulative proportion of how many times a test measured a user's speed above a certain threshold. Now that every unit ID has a proportion representing the proportion of time that a unit ID's tests were measured at a given threshold, a cutoff must be established for the consistent speed metric. The IAA team decided that, counting back from Performance100, the first performance metric that was equal to greater than 0.9 (90%) for a given unit ID represented the consistent speed for that user.

There is no specific industry-wide guideline regarding what percentage of the time an ISP must meet advertised speed, so the team at IAA arbitrarily chose 90% of the time because it seems reasonable that if someone gets a certain speed 90% of the time, it's valid to say they consistently get that speed; the number can be changed to any value deemed reasonable by the FCC. This also allows for some wiggle room when unforeseen technical issues arise that might adversely impact an ISP's score but are beyond their control. In addition, the aggregation can just as easily be performed on an ISP, technology, state (assuming adequate sample sizes), or region level. Unit ID was chosen as it provides the maximum granularity and could be further aggregated by calculating a percentile (10th in this case) or averaging.

Given the fact that this metric accounts for the variation in performance over time, the metric represents not only consistency, but variance as well. This alleviates the need to create more plots and graphs for variance over time because the metric and the confidence

intervals around it incorporate the variation in speed seen throughout the period of observation.

For the code that calculates the metric and the all datasets used for the analyses herein, please refer to the soft appendix Clean Code for Report.sas

4.4.2 Confidence Intervals

The 99.99% Confidence intervals for all ISPs in all states without accounting for peak time.

For the same tables with peak time performance, please refer to the soft appendix All Bootstrapped CIs.xlsx.

4.5 Appendix C (Variable Importance)

This section will explain the methodology of sequential R-Square, decision tree, and linear regression variable selection techniques as well as the scoring system that was developed to rank the final list of the selected explanatory variables. Multiple variable selection techniques were used to minimize bias that a single technique would introduce. The sequential R-square selection technique was executed using the Variable Selection node in SAS Enterprise Miner 7.1. Decision tree and linear regression modeling were also executed using SAS Enterprise Miner 7.1.

Interaction variables were created out of main effects variables to help explain more variability in the data. Interaction variables involve three or more variables, one being the predicted variable and the others being a combination of the explanatory variables. Interaction variables measure how one explanatory variable affects the relationship between the predicted variable and another explanatory variable across different levels. Examples of interaction variables used were Peak x Download Tier or ISP x State x Technology. As one can see from the final ranking, some of the highest ranked variables were interaction variables because they were more statistically significant or were better able to explain variability in the data. A data dictionary explaining every variable in the study is provided in appendix A.

The highest performing models were only able to explain roughly 40% of the variability in the predicted variable. Although there is no standard on how much variability a model in the study should explain, there can be room for improvement. The explanatory power of the models can be increased by looking at more or different variables in future studies on top of the existing variables that were evaluated in this study.

4.5.1 Appendix C1: Variable Selection Node in Enterprise Miner

The Variable Selection node in Enterprise Miner was used to select the best explanatory variables in predicting consistent speed. Primarily the Variable Selection node focuses on selecting variables that can explain a lot of variability in the predicted variable by evaluating the variable's R-square statistic in sequential order. R-square measures how much variability in the predicted variable is explained by the explanatory variable. Variables that can explain a lot of the variability of the predicted variable have high R-square values. To ensure a higher accuracy of selecting the correct variables, the Variable Selection node excludes observations with missing consistent speed values. For missing explanatory class variable values, the variable selection node will create a "missing values" category within the class variable. For interval explanatory variables, missing values will be imputed with the weighted mean of the variable. As previously discussed, the available variables in the data do not do a good job of explaining all the variability in the data. Therefore, the variable selection node compensates for low R-square values by grouping statistically significant levels within class variables. By grouping many different levels of

class variables, the R-square values of the variables increase. (Refer Soft Appendix for excel file called Variable Selection Grouped Levels.xlsx and SAS file Variable Selection Node Code.sas.)

4.5.2 Appendix C2: Decision Tree

The second technique that was used to help determine which variables were most important in explaining consistent speed was decision tree modeling. A decision tree represents a segmentation of the data by applying a series of rules. Each rule assigns an observation to a segment based on the value of one variable. One rule is applied after another, resulting in a hierarchy of data segments. The whole hierarchy is called a tree, and each split point is called a node. The original segment contains the entire data set and is called the root node of the tree. A node with all its successors forms a branch of the node that created it. The final nodes are called leaves. For each leaf, a decision is made and applied to all observations in the leaf. In predictive modeling, the decision is the desired predicted value. Decision trees help determine which variables are most important in explaining the predicted variable because the higher up that segment node is in the tree hierarchy, and the more that variable is used to segment the data the more important it is. Of all the four techniques, Decision Tree modeling was most selective by excluding the most variables. This is in part due to the fact that main effects variables of the decision tree model were not used in Decision Tree modeling. However of all the modeling techniques, Decision Tree modeling was the most accurate with the lowest Average Square Error in predicting Consistent Speed. (Refer Soft Appendix for text files called Decision Tree Variable Selection Output.txt and Final Model Selection Output.txt. and SAS files Decision Tree Variable Selection.sas and Final Model Selection Node Code.sas)

4.5.3 Appendix C3: Linear Regression

Linear regression was also leveraged to help determine which variables most significantly explain consistent speed. Linear regression attempts to predict the value of an interval target as a linear function of one or more independent or explanatory variables. Linear regression variable selection uses p-value statistics when selecting variables. A p-value measures whether an explanatory variable's correlation with the predicted variable is statistically significant or random chance. Lower p-values, usually 0.05 or below, mean the correlation relationship is statistically significant. Three flavors of linear regression variable selection techniques were used to help select the variables that are most significant in helping predict consistent speed: backward, forward, and stepwise selection. Backward selection begins with all candidate explanatory variables in the model and removes variables until all explanatory variables meet a minimum significance level in the model. Forward selection begins with no candidate variables in the model and adds variables until no more significant variables can be added. Stepwise regression begins

much like forward selection but may remove variables that are already in the model that may become insignificant due to the addition of other variables.

Linear regression tested all possible two-way interactions for each of the main effects variables. Those two way interactions that were found significant were kept in all other subsequent selection techniques. By leveraging interaction variables, the explanatory power of the models increased significantly. Each technique can select different significant variables depending on the timing of when a variable is added in relation to what other variables are already in the model. For each technique, a variable had to have a p-value of less than 0.05 to be selected. If at any point a variable's p-value was above 0.05, it was excluded from the model. The three selection methods were used to determine if certain variables are consistently selected by each method. A variable that is selected by multiple methods is more than likely a significant variable in explaining consistent speed. Of the three regression methods, backward selection was the most inclusive, where stepwise and forward regression excluded more variables. Dependent on what other variables are in the model, a variable may or may not be significant, which may result in the variable's inclusion or exclusion from the model. (Refer Soft Appendix for text files called Stepwise Regression Variable Selection Output.txt, Forward Regression Variable Selection Output.txt, and Backward Variable Selection Output.txt and SAS files Stepwise Regression Variable Selection Code.sas, Forward Regression Variable Selection Code.sas, and Backward Regression Variable Selection Code.sas)

4.5.4 Appendix C4: Ranking Methodology

Several variable selection methods were used to determine the best explanatory variables that help predict consistent speed. By leveraging several techniques, variable selection bias can be reduced by ensuring that no single technique would include an insignificant variable or exclude a significant variable. A ranking system was then produced where variables that appeared significant in multiple techniques were scored higher than variables that appeared less frequently across multiple techniques. The scoring system works as follows:

Sequential R-square selection using the Variable Selection node in Enterprise Miner 7.1 automatically groups the significant levels of a class variable in order to increase the variable's R-square statistic. When the Variable Selection node groups the significant levels of a class variable, it creates a new variable called a grouped variable that consists of the significant levels within the class variable and rejects the original class variable. For the sake of consistency, the newly created grouped class variables will be treated as the original class variable. Again, details of the significant levels of the grouped class variables are in Appendix 2. The Variable Selection node automatically ranks the highest R-square value of the newly created grouped variables by sequential R-square value from highest to lowest. In order to score these variables, the highest ranking R-square variable received a

score of 20 because the four techniques selected a total 20 variables. The second ranked variable received a score of 19 and so forth until the Variable Selection node stopped selecting variables.

Decision tree modeling also automatically ranks variable importance. Enterprise Miner's algorithm for ranking a variable involves a combination of how high the log-worth value of the variable's segmentation is and how many times the variable was used in segmenting. The higher the log-worth and the more times a variable is used in segmenting the data, the more important the variable is. A similar approach was used in scoring the decision tree variables as the Variable Importance node. Variables that were ranked higher received higher scores. For example, the highest ranked variable received a score of 20. The second highest ranked variable received a score of 19 and so forth. Decision Tree modeling selected the least amount of variables, therefore the scoring system only decreased in value only as far as the number of variables that were selected by the Decision Tree modeling.

In linear regression, three techniques were used to select variables – stepwise, backward, and forward. The differences in each technique were discussed previously. Because stepwise and forward regression are very similar, the techniques ended up selecting the same variables. In order to avoid double scoring the variables that were selected by stepwise and forward variable selection, only the stepwise selected variables were scored. Because stepwise and forward regression add variables in order of lowest p-value or highest statistical significance, the variables that were added first in each selection step were scored higher. For example the first variable to be added received a score of 20. The second variable to be added received a score of 19 and so forth. Lastly, because backward regression begins with all the variables and eliminates the variables with the highest p-value or lowest significance with each successive step, it was not possible to rank any variables that remained in the model. It is not possible to rank which variables were kept first. They were all kept. In order to score the variables selected by backward regression, an equal score of 20 was given to any variable that was selected by backward selection.

If a variable did not appear in a specific technique, it received zero points for that technique. Summing all four techniques, a variable can receive a score as high as 80 or as low as 1. Of the 20 variables, 19 were selected by Backward regression except for Peak x Technology.

Once a score was given to each variable for each selection technique, the variable's scores were summed. The variables' final ranks were given dependent on the variables' final score. For example, Download Tier x Validated by received a final rank of 11. Below is the sample calculation, which would apply for all variables, for Download Tier x Validated by:

- Sequential R-square Variable Selection node: Earned nine 13 because it ranked eighth in this technique

- Decision Tree Variable Importance: Earned zero points because it was not selected by this technique
- Stepwise Linear Regression: Earned 10 points because it was the eleventh variable to enter the model
- Backward Linear Regression: Earned 20 points because it was selected by this technique
- Peak x Download Tier receives a final score of 43 (13+0+10+20)

See Soft Appendix for excel file Variable Importance Scoring and Data Dictionary.xlsx for complete details on the variable scoring system.

4.6 Appendix D (Building the App)

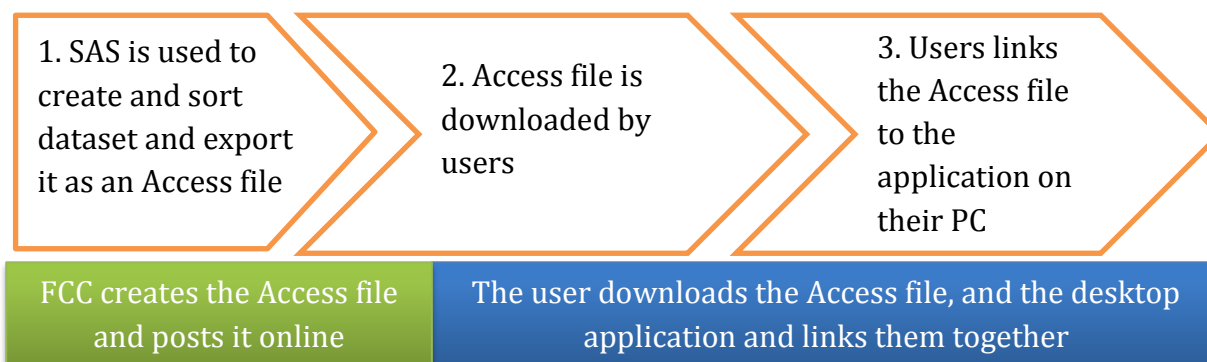
4.6.1 System Requirements for the application

- Microsoft Windows XP or higher to run.
- [dot].NET Framework 3.5¹ or above.
- Visual Basic Power Packs².

4.6.2 Preparing the data

The application reads the data from an Access database and displays it on a desktop window. It is thus critical to ensure that the data is linked the right way for the application to work.

The following steps highlight the procedure to prepare the data for executing the application:



The IAA team created a SAS code (Application_SAS_code.sas), attached in the Soft Appendix, that generates the required Access data file containing all the databases required to use the application.

4.6.3 Executing the application

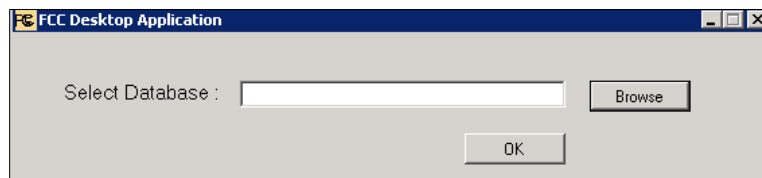
Once the user runs the application, they will be asked to select the location of the Access database created above. The user can then manually link the database like selecting a file using a browse dialog box.

¹ .NET Framework 3.5 can be downloaded at <http://www.microsoft.com/en-us/download/details.aspx?id=21>

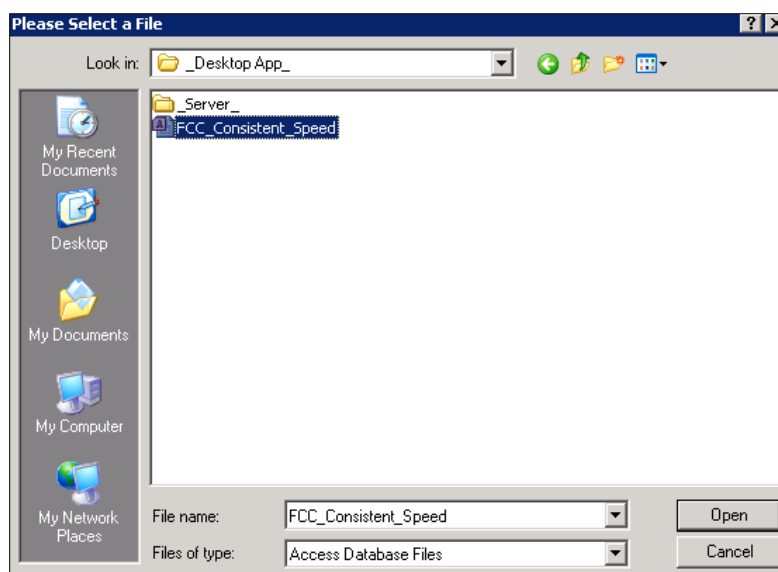
² <http://www.microsoft.com/en-us/download/details.aspx?id=25169>

Following are a set of screenshots that walks through these steps:

1. Run the application



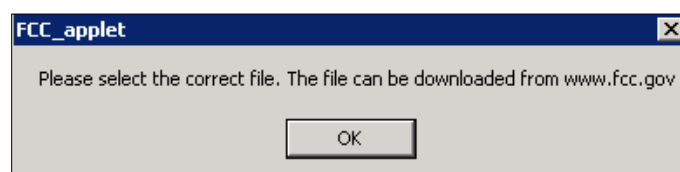
2. Select the Access file location



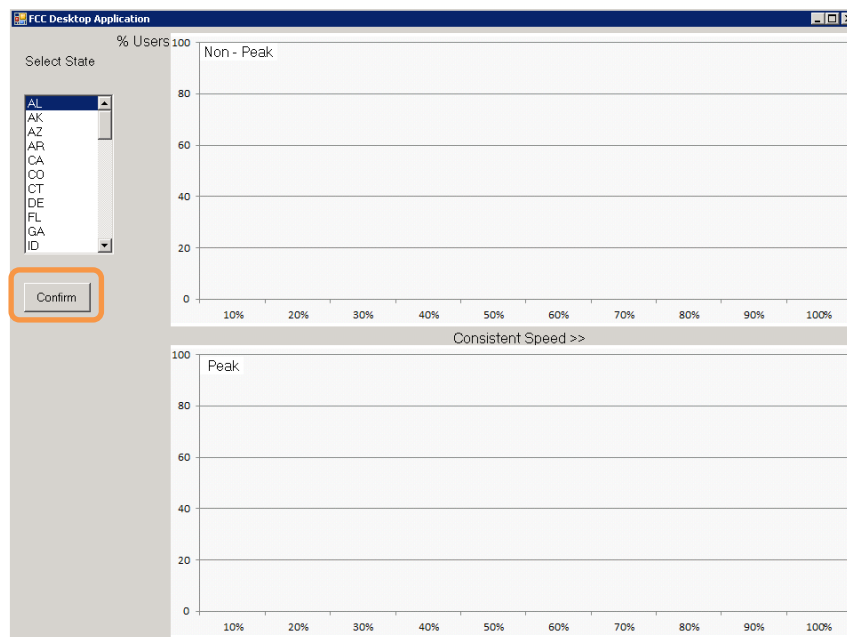
On selecting the correct database one will see the following pop-up saying the file has been successfully linked.



Otherwise one will encounter the following error,

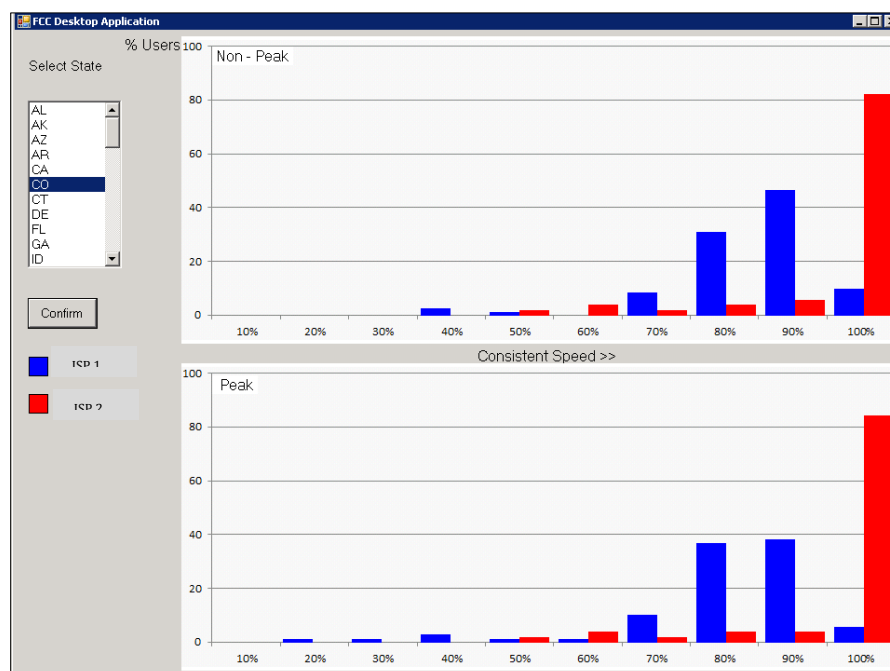


Once the database is linked the user can use the application to view the performance of the top 4 ISPs (or lesser), ranked by the number of unit IDs in his/her state. Once the state is selected, the user should click “Confirm” to see the bar charts.



The first bar chart shows the distribution of the consistent speed metrics in Non-Peak hours, while the second chart shows the same for the peak hours in a day.

The X-axis represents the consistent speed while the Y-axis represents the percentage of total users in the state who experience that level of consistency.

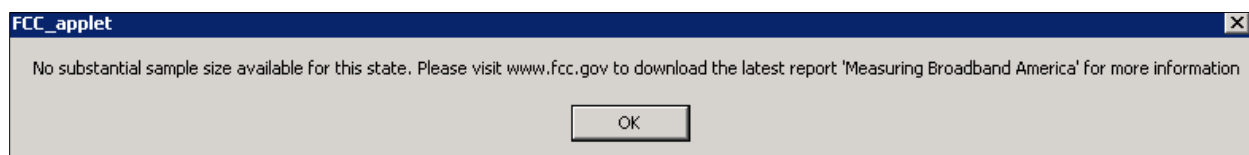


For example, in the above charts, about 45% of Qwest users in Colorado experience a consistent speed between 80% and 90% during non-peak hours, where as that proportion reduces to approximately 38% during peak hours.

4.6.3 Statistical considerations and drawbacks

For statistical considerations the team only took into account those ISPs which had the number of unit IDs > 20. These criteria have been incorporated in the SAS code attached in the technical appendix. Following points touch upon some of the drawbacks of the application:

1. There are quite a few states where sample size was less than 20 unit IDs. On selecting such a state, the user gets the following pop-up message:



In such cases, it is pertinent that the user use the US averages to understand his/her ISP's performance.

2. For states with the number of ISPs less than 4, as per the collected data, the user is able to view the performance of all ISPs in the data set for that state.
3. For states with more than 4 ISPs, the user will be able to compare only the top 4 ISPs by number of unit IDs (sample size)

4.7 Appendix D: Validation and Review

4.7.1 Appendix D1: Methodology

The data-flow and the logic behind the data analysis performed in the existing scripts was understood by the IAA team and the following steps were undertaken:

Step1: The SQL scripts were understood and the logic recoded in SAS. Aggregated data-sets at unit ID level have been generated for multiple tests.

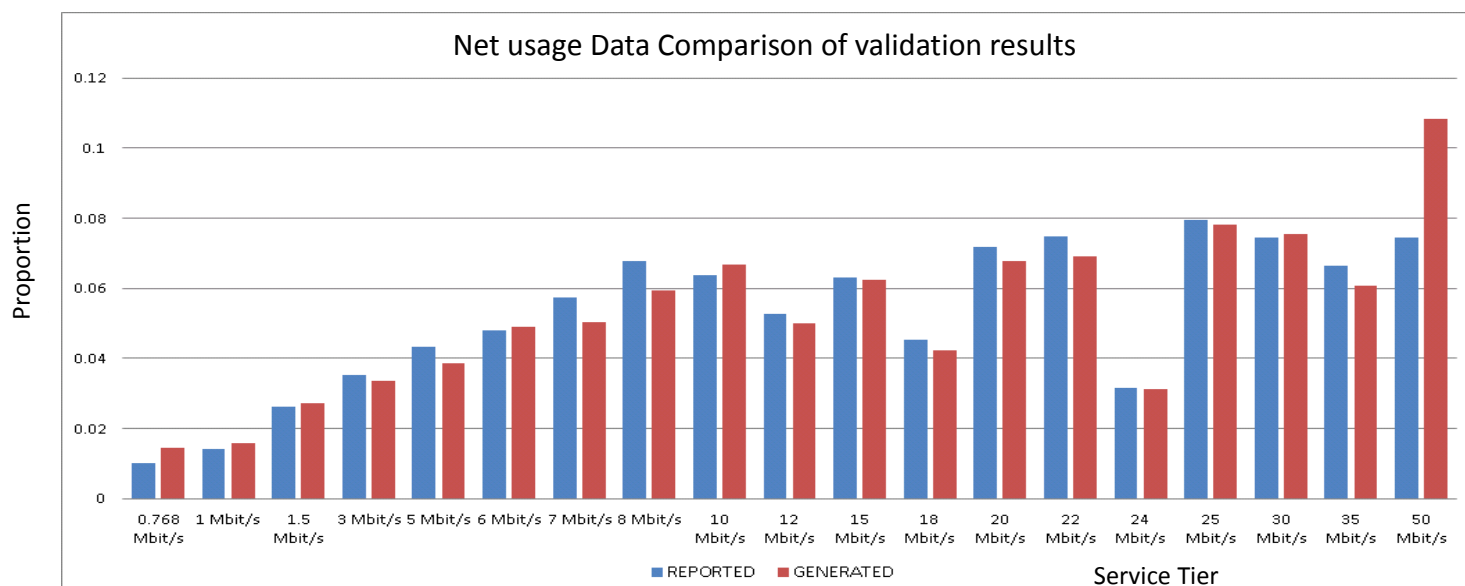
Step2: The SPSS scripts are difficult to understand as the code is poorly commented. Best efforts were put in to extract as much logic as possible e.g. – sample sizes less than twenty five were dropped, Charter in the speed tier 100 was dropped, Brighthouse was dropped etc. In order to proceed with the validation logic was worked out to generate the charts and tables from the unit-level aggregated datasets (produced by Step1). SAS was used for coding this logic. (Refer Soft Appendix for the SAS code for latency test called latency_IAA_sas.doc and netusage_IAA_sas.doc).

Step3: The results produced in Step2 were exported to Excel and tables and charts were prepared. These new tables were compared against the charts in the April report and with the statistical averages excel spread-sheet. (Refer Soft Appendix for the analysis in excel called Validation_Netusage_Chart18.xlsx and Validation_Average_Latency.xlsx).

4.7.2 Appendix D2: Examples of Validation Results

Net-Usage over different speed tiers (Chart #18 in the April Report)


The bars in blue are the ones in the chart on the report, while the red bars are the values generated by using the new scripts. It is seen that the two colored bars closely follow each other. This tells us that the new scripts got results similar to the existing scripts on this test. The only exception is the last speed tier of 50 Mbps. This deviation could be a result of certain data exclusions that were done in the SPSS scripts, which could not be decoded.



4.7.3 Appendix D3: Explanation of Bug

A bug was observed in the existing SQL script and corrected for in the new SAS code. Below is an extract of existing SQL code with the bug highlighted followed by an explanation of the required correction.

```
SELECT unit_id, dtype,
       if(max(fetch_time)-
median(fetch_time)<=3000000,
       if(max(fetch_time)-
min(fetch_time)<=3000000, bytes_sec,
(max(bytes_total)-
min(bytes_total))/((max(fetch_time)-
min(fetch_time))/1000000)),
       (max(bytes_total)-
median(bytes_total))/((max(fetch_time)-
median(fetch_time))/1000000)) as sustained
FROM curr_httpgetmt
INNER JOIN tmp_httpjoin a ON a.d = dtype AND a.u =
```



Explanation of BUG

Such a use of the variable bytes_sec inside a SQL select statement containing a 'group by' SQL statement will cause multiple lines for a single combination of unit_id and dtype. What we want is a single row for every unit_id and dtype and can do this by using an aggregation function like min(bytes_sec), that should have been used instead of bytes_sec.